Document made available under the Patent Cooperation Treaty (PCT)

International application number: PCT/JP05/000461

International filing date: 17 January 2005 (17.01.2005)

Document type: Certified copy of priority document

Document details: Country/Office: JP
Number: 2004-009144
Filing date: 16 January 2004 (16.01.2004)

Date of receipt at the International Bureau: 10 February 2005 (10.02.2005)

Remark: Priority document submitted or transmitted to the International Bureau in compliance with Rule 17.1(a) or (b)

World Intellectual Property Organization (WIPO) - Geneva, Switzerland
Organisation Mondiale de la Propriété Intellectuelle (OMPI) - Genève, Suisse

NEC-1618

PCT/JP2005/000461 PCT

17.1.2005

日本国特許庁
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

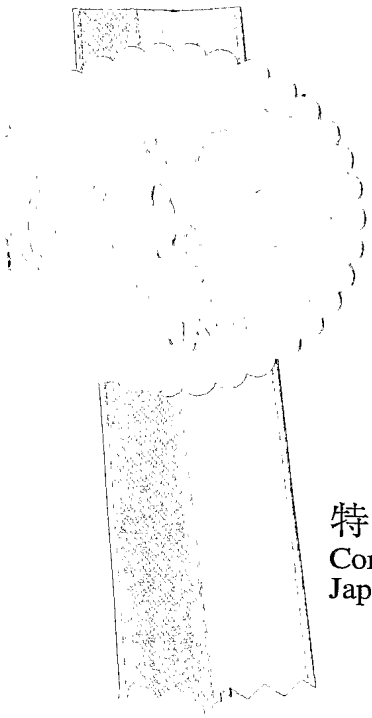
This is to certify that the annexed is a true copy of the following application as filed with this Office.

出願年月日
Date of Application: ２００４年　１月１６日

出願番号
Application Number: 特願２００４－００９１４４

[ST. 10/C]: [ＪＰ２００４－００９１４４]

出願人
Applicant(s): 日本電気株式会社



２００４年　８月３０日

特許庁長官
Commissioner,
Japan Patent Office

小川洋

出証番号　出証特２００４－３０７７５６５

【書類名】　特許願
【整理番号】　34403351
【提出日】　平成16年 1月16日
【あて先】　特許庁長官　　殿
【国際特許分類】　G06F 17/21
【発明者】
　【住所又は居所】　東京都港区芝五丁目７番１号　日本電気株式会社内
　【氏名】　越仲　孝文
【特許出願人】
　【識別番号】　000004237
　【氏名又は名称】　日本電気株式会社
【代理人】
　【識別番号】　100123788
　【弁理士】
　【氏名又は名称】　宮崎　昭夫
　【電話番号】　03-3585-1882
【選任した代理人】
　【識別番号】　100088328
　【弁理士】
　【氏名又は名称】　金田　暢之
【選任した代理人】
　【識別番号】　100106297
　【弁理士】
　【氏名又は名称】　伊藤　克博
【選任した代理人】
　【識別番号】　100106138
　【弁理士】
　【氏名又は名称】　石橋　政幸
【手数料の表示】
　【予納台帳番号】　201087
　【納付金額】　21,000円
【提出物件の目録】
　【物件名】　特許請求の範囲　1
　【物件名】　明細書　1
　【物件名】　図面　1
　【物件名】　要約書　1
　【包括委任状番号】　0304683

【書類名】特許請求の範囲
【請求項１】
コンピュータが、入力されたテキスト文書を話題ごとに分割するテキスト処理方法において、
テキスト文書の話題を隠れ変数に、テキスト文書を観測変数にそれぞれ対応づけた確率モデルを生成する仮モデル生成ステップと、
前記仮モデル生成ステップで生成された確率モデルを規定するモデルパラメータの初期値を出力するモデルパラメータ初期化ステップと、
前記モデルパラメータ初期化ステップで出力されたモデルパラメータの初期値と、入力されたテキスト文書にもとづいて、当該確率モデルを規定するモデルパラメータを推定するモデルパラメータ推定ステップと、
前記モデルパラメータ推定ステップで推定されたモデルパラメータにもとづいて、入力されたテキスト文書を話題ごとに分割するテキスト分割結果出力ステップを有することを特徴とするテキスト処理方法。
【請求項２】
前記仮モデル生成ステップと、前記モデルパラメータ初期化ステップと、前記モデルパラメータ推定ステップにおいて、複数の確率モデルについて、それぞれの処理が行われる場合、前記モデルパラメータ推定ステップで推定されたモデルパラメータにもとづいて、前記複数の確率モデルの中から、前記テキスト分割結果出力ステップで処理を行う確率モデルを選択するモデル選択ステップをさらに有する、請求項１に記載のテキスト処理方法。
【請求項３】
前記確率モデルは隠れマルコフモデルである、請求項１または２に記載のテキスト処理方法。
【請求項４】
前記隠れマルコフモデルは一方向型の構造を有する、請求項３に記載のテキスト処理方法。
【請求項５】
前記隠れマルコフモデルは離散出力型である、請求項３または４に記載のテキスト処理方法。
【請求項６】
前記モデルパラメータ推定ステップでは、最尤推定と、最大事後確率推定と、ベイズ推定のいずれかを用いて、モデルパラメータを推定する、請求項３から５のいずれか１項に記載のテキスト処理方法。
【請求項７】
前記モデル選択ステップでは、赤池情報量基準と、最小記述長基準と、ベイズ事後確率のいずれかを用いて、確率モデルを選択する、請求項３から６のいずれか１項に記載のテキスト処理方法。
【請求項８】
請求項１から７のいずれか１項に記載の各ステップをコンピュータに実行させるプログラム。
【請求項９】
請求項１から７のいずれか１項に記載の各ステップをコンピュータに実行させるプログラムを記録した、コンピュータ読み取りが可能な記録媒体。
【請求項１０】
入力されたテキスト文書を話題ごとに分割するテキスト処理装置において、
テキスト文書の話題を隠れ変数に、テキスト文書を観測変数にそれぞれ対応づけた確率モデルを生成する仮モデル生成手段と、
前記仮モデル生成手段が生成した確率モデルを規定するモデルパラメータの初期値を出力するモデルパラメータ初期化手段と、

前記モデルパラメータ初期化手段によって出力されたモデルパラメータの初期値と、入力されたテキスト文書にもとづいて、当該確率モデルを規定するモデルパラメータを推定するモデルパラメータ推定手段と、

前記モデルパラメータ推定手段が推定したモデルパラメータにもとづいて、入力されたテキスト文書を話題ごとに分割するテキスト分割結果出力手段を有することを特徴とするテキスト処理装置。

【請求項１１】

前記仮モデル生成手段と、前記モデルパラメータ初期化手段と、前記モデルパラメータ推定手段が、複数の確率モデルについて、それぞれの処理を行う場合、前記モデルパラメータ推定手段によって推定されたモデルパラメータにもとづいて、前記複数の確率モデルから１つの確率モデルを選択し、前記テキスト分割結果出力手段に対して、当該確率モデルについて処理を行わせるモデル選択手段をさらに有する、請求項１０に記載のテキスト処理装置。

【請求項１２】

前記確率モデルは隠れマルコフモデルである、請求項１０または１１に記載のテキスト処理装置。

【請求項１３】

前記隠れマルコフモデルは一方向型の構造を有する、請求項１２に記載のテキスト処理装置。

【請求項１４】

前記隠れマルコフモデルは離散出力型である、請求項１２または１３に記載のテキスト処理装置。

【請求項１５】

前記モデルパラメータ推定手段は、最尤推定と、最大事後確率推定と、ベイズ推定のいずれかを用いて、モデルパラメータを推定する、請求項１２から１４のいずれか１項に記載のテキスト処理装置。

【請求項１６】

前記モデル選択手段は、赤池情報量基準と、最小記述長基準と、ベイズ事後確率のいずれかを用いて、確率モデルを選択する、請求項１２から１５のいずれか１項に記載のテキスト処理装置。

【書類名】明細書
【発明の名称】テキスト処理方法／プログラム／プログラム記録媒体／装置
【技術分野】
【０００１】
本発明は、文字列や単語列といったテキスト文書を、意味的にまとまった部分ごとに、すなわち話題ごとに分割するテキスト処理方法／プログラム／プログラム記録媒体／装置に関する。
【背景技術】
【０００２】
この種のテキスト処理方法／プログラム／プログラム記録媒体／装置は、長大かつ多数のテキスト文書を意味内容ごとに、すなわち話題ごとに分割、分類等することによって、人がテキスト文書から所望の情報を得やすいように加工することを目的として用いられている。ここでテキスト文書とは、例えば、磁気ディスク等の記録媒体に記録された任意の文字や単語などの並びである。あるいは、紙に印刷されたり、タブレットに手書きされたりした文字列を光学的文字読取り装置(OCR)で読み取った結果や、人の発話で生じる音声波形信号を音声認識装置で認識した結果等も、テキスト文書である。さらに一般的には、毎日の天候の記録、店舗における商品の販売記録、コンピュータを操作した際のコマンドの記録、等々、時系列的に生成される記号の並びのほとんどは、テキスト文書の範疇に入る。
【０００３】
この種のテキスト処理方法／プログラム／プログラム記録媒体／装置に関して、大別して2種類の従来技術が挙げられる。これら2種類の従来技術について、図面を参照して詳細に説明する。
【０００４】
第１の従来技術は、入力テキストを単語の系列o_1，o_2，…，o_Tとして、系列中の各区間で単語の出現傾向に関する統計量を算出し、この統計量に急激な変化がみられる位置を話題の変化点として検出する。図5は、第１の従来技術の比較的単純な例を概念的に表している。図5に示すように、入力テキストの各部分に対して一定幅の窓を設定し、窓内における単語の出現回数を計数し、単語の出現頻度を多項分布の形式で算出する。そして、近接する2つの窓(図5における窓1および窓2)の間の差異が所定のしきい値より大きければ、これら2つの窓の境界で話題の変化が起こったと判定する。2窓間の差異には、例えば[数１]で表されるような、窓ごとに計算された多項分布間のKLダイバージェンスを用いることができる。
【０００５】
【数１】

$$\sum_{i=1}^{L} a_i \log \frac{a_i}{b_i}$$

ここで、a_i，b_i（i=1，…，L）はそれぞれ窓1、窓2に対応する単語の出現頻度を表す多項分布で、$a_1+a_2+\cdots+a_L=1$，$b_1+b_2+\cdots+b_L=1$ を満たす。Lは入力テキストの語彙数である。
【０００６】
上では特に、窓内の統計量を個々の単語の出現頻度から計算する、いわゆるユニグラム(unigram)としているが、隣接2つ組、3つ組、さらには任意個の組の単語出現頻度(それぞれバイグラムbigram、トライグラムtrigram、n-gram)を考えてもよい。あるいは、文献「2001年11月、情報処理学会論文誌、第42巻、第11号、第2650～2662頁、別所克人、単語の概念ベクトルを用いたテキストセグメンテーション」（非特許文献１）に記載されているように、隣接しない単語同士の共起を考慮することにより、入力テキスト中の各単語を実ベクトルに置き換えて、このベクトルの移動量の多さで話題の変化点を検出することも

できる。

【０００７】

第２の従来技術は、種々の話題に関する統計的モデルをあらかじめ準備しておき、それらのモデルと入力単語列の最適なマッチングを計算することにより、話題の推移を求める。第２の従来技術の例は、「2000年、プロシーディング・オブ・フォース・ユーロピアン・カンファレンス・オン・リサーチ・アンド・アドバンスト・テクノロジ・フォー・ディジタル・ライブラリ、アマラル他、トピック・ディテクション・イン・レッド・ドキュメント（Amaral et al., Topic Detection in Read Documents, Proceedings of 4th European Conference on Research and Advanced Technology for Digital Libraries, 2000）」（非特許文献２）に記載されている。図6は、非特許文献２に記載されている第２の従来技術の例を概念的に示したものである。この第２の従来技術の例は、図6に示すように、「政治」、「スポーツ」、「経済」などといった話題ごとに、話題ごとの統計モデル、つまり話題モデルを作成して準備しておく。話題モデルは、あらかじめ話題ごとに大量収集されたテキスト文書から求めた単語出現頻度(ユニグラム、バイグラム等)である。このように話題モデルを準備し、これら話題間の遷移の起こりやすさ(遷移確率)を適宜決めておけば、入力単語系列ともっともよく整合する話題モデル系列を機械的に算出することができる。仮に、入力単語系列を入力音声波形と置き換えて、話題モデルを音素モデルに置き換えてみれば容易にわかるように、音声認識に関して多数ある従来技術と同様に、DPマッチングの要領で、フレーム同期ビームサーチなどの計算法を利用して話題の遷移系列を計算することができる。

【０００８】

上で述べた第２の従来技術の例は、「政治」、「スポーツ」、「経済」など、人間が直感的に理解しやすい話題を設定して、話題の統計モデルを作成しているが、「1998年、プロシーディング・オブ・インターナショナル・カンファレンス・オン・アクースティック・スピーチ・アンド・シグナル・プロセッシング98、第1巻、333～336頁、ヤムロン他、ヒドゥン・マルコフ・モデル・アプローチ・トゥ・テキスト・セグメンテーション・アンド・イベント・トラッキング（Yamron et al., Hidden Markov model approach to text Segmentation and event tracking, Proceedings of International Conference on Acoustic, Speech and Signal Processing 98, Vol.1, pp.333-336, 1998)」（非特許文献３）に記載があるように、テキスト文書に対して何らかの自動クラスタリング手法を適用して、人間の直感とは無関係な話題モデルを作る例もある。この場合、話題モデルを作るために大量のテキスト文書を話題ごとに分類しておく必要がないので、手間は幾分少なくてすむ。ただし、大規模なテキスト文書集合を用意して、そこから話題モデルを作成するという点は同様である。

【非特許文献１】「2001年11月、情報処理学会論文誌、第42巻、第11号、第2650～2662頁、別所克人、単語の概念ベクトルを用いたテキストセグメンテーション」

【非特許文献２】「2000年、プロシーディング・オブ・フォース・ユーロピアン・カンファレンス・オン・リサーチ・アンド・アドバンスト・テクノロジ・フォー・ディジタル・ライブラリ、アマラル他、トピック・ディテクション・イン・レッド・ドキュメント（Amaral et al., Topic Detection in Read Documents, Proceedings of 4th European Conference on Research and Advanced Technology for Digital Libraries, 2000)」

【非特許文献３】「1998年、プロシーディング・オブ・インターナショナル・カンファレンス・オン・アクースティック・スピーチ・アンド・シグナル・プロセッシング98、第1巻、333～336頁、ヤムロン他、ヒドゥン・マルコフ・モデル・アプローチ・トゥ・テキスト・セグメンテーション・アンド・イベント・トラッキング（Yamron et al., Hidden Markov model approach to text Segmentation and event tracking, Proceedings of International Conference on Acoustic, Speech and Signal Processing 98, Vol.1, pp.333-336, 1998)」

【非特許文献４】「1995年11月、NTTアドバンステクノロジ株式会社、ラビナー他著

、古井他訳、音声認識の基礎(下)、第129～134頁」
【非特許文献５】「1994年12月、岩波書店、岩波講座応用数学[対象11]、韓太舜他著、情報と符号化の数理、第249～275頁」
【非特許文献６】「1995年11月、NTTアドバンステクノロジ株式会社、ラビナー他著、古井他訳、音声認識の基礎(下)、第166～169頁」
【非特許文献７】「1995年11月、NTTアドバンステクノロジ株式会社、ラビナー他著、古井他訳、音声認識の基礎(下)、第280～281頁」
【非特許文献８】「2002年７月、電子情報通信学会誌、第85巻、第7号、第504～509頁、上田、ベイズ学習〔III〕 ―変分ベイズ学習の基礎―」

【発明の開示】
【発明が解決しようとする課題】
【０００９】
しかしながら、上述した第１の従来技術および第２の従来技術は、それぞれいくつかの問題を有する。
【００１０】
第１の従来技術では、窓間の差異に関するしきい値や、窓幅といったパラメータを最適に調整することが難しいという問題がある。あるテキスト文書に対して所望の分割がなされるようにパラメータ値を調整することは、可能な場合もある。しかし、そのために試行錯誤的にパラメータ値を調整する手間が必要である。加えて、仮にあるテキスト文書に対して所望の動作が実現できたとしても、同じパラメータ値を別のテキスト文書に適用した場合、期待通りに動作しないことが多い。なぜなら、例えば窓幅というパラメータは、大きくすればするほど窓内の単語出現頻度を正確に見積もることができるから、テキストの分割処理も正確に実行できるが、窓幅は入力テキスト中の話題の長さよりも長いと、明らかに話題分割という当初の目的を達せられなくなる。すなわち、入力テキストの性質によって、窓幅の最適値は異なる。窓間の差異に関するしきい値も同様で、入力テキストに応じてその最適値が異なるのが普通である。これは、入力テキスト文書の性質によっては期待通りの動作をしないということであるから、実際応用上深刻な問題となる。
【００１１】
第２の従来技術では、話題のモデルを作成するために、事前に大規模なテキストコーパスを準備しなければならないという問題がある。しかもそのテキストコーパスは、話題ごとに分割済みであることが必須であり、しばしば話題のラベル(例えば「政治」、「スポーツ」、「経済」等)が付与されていることが要求される。このようなテキストコーパスを事前に準備するのには、当然時間と費用がかかる。しかも、第２の従来技術では、話題のモデルを作成するのに使用したテキストコーパスが、入力テキスト中の話題と同じ話題を含んでいること、すなわちドメインが一致していることが必要となる。したがって、この従来技術の例の場合、入力テキストのドメインが未知の場合、またはドメインが頻繁に変化し得る場合、所望のテキスト分割結果を得ることは困難である。
【００１２】
そこで本発明の目的は、入力テキスト文書の性質によってパラメータを調整する手間が不要で、事前に時間と費用をかけて大規模なテキストコーパスを準備する必要もなく、なおかつ入力テキストのドメインに依存せずにテキストを話題ごとに分割できるテキスト処理方法／プログラム／プログラム記録媒体／装置を提供することにある。
【課題を解決するための手段】
【００１３】
上記目的を達成するために、本発明は、入力されたテキスト文書を生成したと推測されるテキスト文書の話題を隠れ変数に、テキスト文書を観測変数にそれぞれ対応付けた、１つまたは複数の確率モデルを生成し、入力されたテキスト文書にもとづいて、生成された確率モデルを規定するモデルパラメータを推定する。モデルパラメータが推定された後、この推定結果にもとづいて、入力されたテキスト文書中の話題の推移を確率的に算出する。

【００１４】
入力されたテキスト文書に含まれる話題数について複数の可能性を想定し、複数の確率モデルを生成した場合、最良の確率モデルを選択し、この最良の確率モデルに対応するパラメータ推定結果にもとづいて、入力されたテキスト文書中の話題の推移を算出する。
【発明の効果】
【００１５】
以上説明したように、本発明によれば、入力テキスト文書の性質に応じてパラメータを調整する手間が少なく、事前に時間と費用をかけて大規模なテキストコーパスを準備する必要もなく、なおかつ入力テキストがどのような内容を含んでいるか、すなわちドメインに依存せずに、テキストを精度よく話題ごとに分割することが可能となる。
【発明を実施するための最良の形態】
【００１６】
（第１の実施形態）
次に、本発明の第１の実施形態について、図面を参照して詳細に説明する。
【００１７】
図1は、本発明の第１の実施形態のテキスト処理装置の構成を示すブロック図である。本実施形態のテキスト処理装置は、テキスト文書を入力するテキスト入力部101と、入力されたテキスト文書を格納するテキスト記憶部102と、テキスト文書の話題の推移を記述する、テキスト文書の話題を隠れ変数に、テキスト文書を観測変数にそれぞれ対応付けた、単一もしくは複数のモデルを生成する仮モデル生成部103と、仮モデル生成部103が生成した各モデルを規定する各モデルパラメータの値を初期化するモデルパラメータ初期化部104と、モデルパラメータ初期化部104によって初期化されたモデルとテキスト記憶部102に格納されたテキスト文書を使ってモデルパラメータを推定するモデルパラメータ推定部105と、モデルパラメータ推定部105が行ったパラメータ推定の結果を格納する推定結果記憶部106と、推定結果記憶部106に複数のモデルのパラメータ推定結果が格納されている場合にその中から1つのモデルのパラメータ推定結果を選択するモデル選択部107と、モデル選択部107が選択したモデルのパラメータ推定結果から入力テキスト文書の分割を行って結果を出力するテキスト分割結果出力部108を備える。各々の部は、それぞれ計算機上に記憶されたプログラムによって、またはこのプログラムが記録された記録媒体を読み取ることによって動作させることにより実現可能である。
【００１８】
ここでテキスト文書とは、上述したように、例えば、磁気ディスク等の記録媒体に記録された任意の文字や単語などの並びである。あるいは、紙に印刷されたりタブレットに手書きされたりした文字列を光学的文字読取り装置(OCR)で読み取った結果や、人の発話で生じる音声波形信号を音声認識装置で認識した結果等も、テキスト文書である。さらに一般的には、毎日の天候の記録、店舗における商品の販売記録、コンピュータを操作した際のコマンドの記録、等々、時系列的に生成される記号の並びのほとんどは、テキスト文書の範疇に入る。
【００１９】
次に、本実施形態のテキスト処理装置の動作を、図２を参照して詳細に説明する。
【００２０】
テキスト入力部101から入力されたテキスト文書は、テキスト記憶部102に格納される（ステップ２０１）。ここでテキスト文書は、多数、例えばT個の単語が一列に並んだ単語系列とし、以下では $o_1, o_2, \cdots, o_T$ と表すことにする。単語間にスペースのない日本語の場合は、テキスト文書に対して公知の形態素解析法を適用することにより、単語に分割すればよい。また、この単語列から、テキスト文書の話題とは直接関係のない助詞や助動詞などをあらかじめ取り除いて、名詞や動詞などの重要語のみの単語列としてもよい。これには、公知の形態素解析法によって各単語の品詞を求め、名詞、動詞、形容詞などを重要語として取り出すようにすればよい。さらには、入力テキスト文書が、音声信号を音声認識して得られた音声認識結果であり、かつ音声信号に一定時間以上継続する無音(発話

休止)区間が存在する場合は、テキスト文書の対応する位置に<ポーズ>のような単語を含めてよい。同様に、入力テキスト文書が、紙文書をOCRにかけることによって得られた文字認識結果である場合には、<改行>のような単語をテキスト文書中の対応する位置に含めてよい。

【００２１】

なお、通常の意味での単語系列(ユニグラム，unigram)の代わりに、隣接する単語の2つ組(バイグラム，bigram)、3つ組(トライグラム，trigram)、さらに一般的なn個組(n-gram)を一種の単語と考えて、その系列をテキスト記憶部102に格納してもよい。例えば2つ組での単語列の格納形式は（o_1，o_2)，(o_2，o_3)，…，(o_{T-1}，o_T）となり、系列の長さはT-1である。

【００２２】

仮モデル生成部103は、入力されたテキスト文書を生成したと推測される単一もしくは複数の確率モデルを生成する。ここで確率モデルまたはモデルとは、一般にはグラフィカルモデルと呼ばれる、複数のノードとそれらを結ぶアークとで表現されるモデル全般を指す。グラフィカルモデルには、マルコフモデルやニューラルネットワーク、ベイジアンネットなどが含まれる。本実施形態においては、ノードがテキスト中に含まれる話題に対応する。また、モデルから生成されて観測される観測変数には、テキスト文書の構成要素であるところの単語が対応する。

【００２３】

本実施形態では、モデルを隠れマルコフモデル(Hidden Markov ModelまたはHMM)とし、なおかつその構造は一方向型(left-to-right型)で、出力は上述の入力単語列に含まれる単語の系列(離散値)とする。Left-to-right型HMMでは、ノードの数を指定すればモデルの構造が一意に決定される。このモデルの概念図を図3に示す。HMMの場合特に、ノードのことを状態と呼ぶのが一般的である。図3の場合、ノード数、すなわち状態数は4である。

【００２４】

仮モデル生成部103は、入力テキスト文書にいくつの話題が含まれているかに応じて、モデルの状態数を決定し、その状態数に応じてモデルすなわちHMMを生成する。例えば、入力テキスト文書に4個の話題が含まれているとわかっていれば、仮モデル生成部103は4状態のHMMを1つだけ生成する。また、入力テキスト文書に含まれる話題の数が未知の場合は、十分小さい状態数N_{min}のHMMから、十分大きい状態数N_{max}のHMMまでのすべての状態数のHMMを、各々1つずつ生成する（ステップ２０２、２０６、２０７）。ここでモデルを生成するとは、モデルを規定するパラメータの値を記憶するための記憶領域を記憶媒体上に確保する、という意味である。モデルを規定するパラメータについては後述する。

【００２５】

モデルパラメータ初期化部104は、仮モデル生成部103が生成したすべてのモデルについて、モデルを規定するパラメータの値を初期化する（ステップ２０３）。モデルを規定するパラメータは、上述のleft-to-right型離散HMMの場合、状態遷移確率 a_1，a_2，…，a_N、および記号出力確率$b_{1,j}$，$b_{2,j}$，…，$b_{N,j}$ とする。ここにNは状態数である。また j=1，2，…，L で、Lは入力テキスト文書に含まれる単語の種類数、すなわち語彙数である。状態遷移確率a_iは、状態iから状態i+1に遷移する確率であり、$0<a_i \leq 1$でなければならない。よって、状態iから再度状態iに戻る確率は$1-a_i$となる。また、記号出力確率$b_{i,j}$は、ある一度の状態遷移の後に、状態iに至ったとして、インデクスjで指定される単語が出力される確率である。すべての状態i=1，2，…，Nにおいて、記号出力確率の総和$b_{i,1}+b_{i,2}+\cdots+b_{i,L}$は1でなければならない。

【００２６】

モデルパラメータ初期化部104は、状態数Nのモデルに対して、例えば上述の各パラメータの値を$a_i=N/T$、$b_{i,j}=1/L$のように設定する。この初期値の与え方に決まったやり方はなく、上述の確率の条件さえ満たしていれば、いろいろな方法があり得る。ここで述べた方法はほんの一例である。

【００２７】

モデルパラメータ推定部105は、モデルパラメータ初期化部104によって初期化された単一もしくは複数のモデルを順次受け取り、モデルが入力テキスト文書o_1，o_2，…，o_Tを生成する確率、すなわち尤度がなるべく高くなるように、モデルパラメータを推定する（ステップ２０４）。これには公知の最尤推定法、特に、反復計算を基本とする期待値最大化法(EM法)を用いることができる。すなわち、例えば文献「1995年11月、NTTアドバンステクノロジ株式会社、ラビナー他著、古井他訳、音声認識の基礎(下)、第129～134頁」（非特許文献４）に記載されているように、その時点で得られているパラメータ値a_i、$b_{i,j}$を用いて、[数２]に示す漸化式によって前向き変数$\alpha_t(i)$および後向き変数$\beta_t(i)$をt=1，2，…，T、i=1，2，…，Nにわたって計算し、さらに[数３]に従ってパラメータ値を再計算する。再計算されたパラメータ値を用いて再度[数２]および[数３]を計算する。以下、収束するまで十分な回数これをくり返す。ただしここにδ_{ij}はクロネッカーのデルタ、すなわち、i=jなら1、そうでなければ0をとる。

【００２８】

【数２】

$$\alpha_1(i)=b_{1,o_1}\delta_{1,i},\ \alpha_t(i)=a_{i-1}b_{i,o_t}\alpha_{t-1}(i-1)+(1-a_i)b_{i,o_t}\alpha_{t-1}(i),$$
$$\beta_T(i)=a_N\delta_{N,i},\ \beta_t(i)=(1-a_i)b_{i,o_{t+1}}\beta_{t+1}(i)+a_ib_{i+1,o_{t+1}}\beta_{t+1}(i+1).$$

【００２９】

【数３】

$$a_i \leftarrow \frac{\sum_{t=1}^{T-1}\alpha_t(i)a_ib_{i+1,o_t}\beta_{t+1}(i+1)}{\sum_{t=1}^{T-1}\alpha_t(i)(1-a_i)b_{i,o_t}\beta_{t+1}(i)+\sum_{t=1}^{T-1}\alpha_t(i)a_ib_{i+1,o_t}\beta_{t+1}(i+1)},$$

$$b_{ij} \leftarrow \frac{\sum_{t=1}^{T}\alpha_t(i)\beta_t(i)\delta_{j,o_t}}{\sum_{t=1}^{T}\alpha_t(i)\beta_t(i)}.$$

【００３０】

モデルパラメータ推定部105におけるパラメータ推定の反復計算の収束判定を行うには、尤度の上昇量をみればよい。すなわち、上述の反復計算によって尤度の上昇がみられなくなれば、その時点で反復計算を終了すればよい。ここで、尤度は$\alpha_1(1)\beta_1(1)$として得られる。モデルパラメータ推定部105は、反復計算を終了した時点で、モデルパラメータa_i、$b_{i,j}$と、前向きおよび後向き変数$\alpha_t(i)$、$\beta_t(i)$を、モデル(HMM)の状態数と対にして、推定結果記憶部106に格納する（ステップ２０５）。

【００３１】

モデル選択部107は、モデルパラメータ推定部105で状態数ごとに得られたパラメータ推定結果を推定結果記憶部106から受け取り、各モデルの確からしさを計算し、もっとも確からしいモデルを1つ選択する（ステップ２０８）。モデルの確からしさは、公知の赤池情報量基準(AIC)や最小記述長基準(MDL基準)などに基づいて計算することができる。赤池情報量基準、最小記述長基準については、例えば文献「1994年12月、岩波書店、岩波講座応用数学[対象11]、韓太舜他著、情報と符号化の数理、第249～275頁」（非特許文献５）に記載がある。例えばAICによれば、パラメータ推定収束後の対数尤度 $\log(\alpha_1(1)\beta_1(1)$

）とモデルパラメータ数NLの差が最大となるモデルが選択される。また、MDLによれば、近似的に、対数尤度を符号反転した $-\log(\alpha_1(1)\beta_1(1))$ と、モデルパラメータ数と入力テキスト文書の単語系列長の平方根との積NL×log(T)/2の和が最小となるモデルが選択される。なお、AICでもMDLでも、モデルパラメータ数NLに関わる項に、経験的に決まる定数係数をかけて、選択されるモデルを意図的に調整する操作が一般的に行われているが、本実施形態でもそのような操作は行って差し支えない。

【００３２】

テキスト分割結果出力部108は、モデル選択部107によって選択された状態数Nのモデルに対応するモデルパラメータ推定結果を推定結果記憶部106から受け取り、この推定結果における入力テキスト文書に対する話題ごとの分割結果を算出する（ステップ２０９）。状態数Nのモデルによる分割は、入力テキスト文書o_1，o_2，…，o_TをN個の区間に分割する。分割結果は、まず[数４]に従って、確率的に計算される。[数４]は、入力テキスト文書中の単語o_tが第i番目の話題区間に割り当てられる確率を示す。最終的な分割結果は、$P(z_t=i \mid o_1, o_2, \cdots, o_T)$ が最大となるiをt=1，2，…，T にわたって求めることで得られる。

【００３３】

【数４】

$$P(z_t = i \mid o_1, o_2, \cdots, o_T) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{N}\alpha_t(j)\beta_t(j)}$$

【００３４】

なお、ここではモデルパラメータ推定部105は、最尤推定法を用いて、すなわち[数３]を用いて、パラメータを逐次更新したが、最尤推定法の他に、最大事後確率推定(MAP推定)を用いることもできる。最大事後確率推定については、例えば文献「1995年11月、NTTアドバンステクノロジ株式会社、ラビナー他著、古井他訳、音声認識の基礎(下)、第166～169頁」（非特許文献６）に記載がある。最大事後確率推定の場合、例えばモデルパラメータの事前分布に共役事前分布を用いると、a_iの事前分布はベータ分布 $\log p(a_i \mid \kappa_0, \kappa_1) = (\kappa_0-1)\times\log(1-a_i) + (\kappa_1-1)\times\log(a_i) + \text{const}$、$b_{ij}$の分布はディレクレ分布 $\log p(b_{i,1}, b_{i,2}, \cdots, b_{i,L} \mid \lambda_1, \lambda_2, \cdots, \lambda_L) = (\lambda_1-1)\times\log(b_{i,1}) + (\lambda_2-1)\times\log(b_{i,2}) + \cdots + (\lambda_L-1)\times\log(b_{i,L}) + \text{const}$ と表される。ただしκ_0，κ_1，λ_1，λ_2，…，λ_Lおよびconstは定数である。このとき、最尤推定の[数３]に相当する最大事後確率推定のパラメータ更新式は、[数５]のように表される。

【００３５】

【数５】

$$a_i \leftarrow \frac{\sum_{t=1}^{T-1}\alpha_t(i)a_i b_{i+1,o_t}\beta_{t+1}(i+1)+\kappa_1-1}{\sum_{t=1}^{T-1}\alpha_t(i)(1-a_i)b_{i,o_t}\beta_{t+1}(i)+\kappa_0-1+\sum_{t=1}^{T-1}\alpha_t(i)a_i b_{i+1,o_t}\beta_{t+1}(i+1)+\kappa_1-1},$$

$$b_{ij} \leftarrow \frac{\sum_{t=1}^{T}\alpha_t(i)\beta_t(i)\delta_{j,o_t}+\lambda_j-1}{\sum_{t=1}^{T}\alpha_t(i)\beta_t(i)+\sum_{k=1}^{L}(\lambda_k-1)}.$$

【００３６】

なお、ここまでで述べた本実施形態においては、記号出力確率b_{ij}が状態と対応付けられている。すなわち、単語がHMMの各状態(ノード)から発生するとするモデルを用いている。しかし、単語が状態遷移(アーク)から発生するとするモデルを用いることも可能である。例えば入力テキストが紙文書のOCR結果であったり、音声信号の音声認識結果であったりする場合、単語が状態遷移から発生するようなモデルは便利である。なぜなら、音声信号における発話休止や、紙文書における改行などを意味する単語、すなわち<ポーズ>や<改行>などが含まれたテキスト文書の場合は、状態iからi+1への状態遷移から発生する単語が必ず<ポーズ>や<改行>であるように、記号出力確率を固定しておけば、本実施形態によって入力テキスト文書から検出される話題境界には、必ず<ポーズ>や<改行>が当てはまるようにできる。また、仮に入力テキスト文書がOCR結果や音声認識結果ではなくとも、単語が状態遷移から発生するモデルで、状態iからi+1への状態遷移から、「では」、「次に」、「さて」などといった、話題の切り替わりと関連の深い単語が発生するように記号出力確率を設定しておけば、検出される話題境界には「では」、「次に」、「さて」などの単語が現れやすくできる。

【００３７】

（第２の実施形態）

次に、本発明の第２の実施形態について、図面を参照して詳細に説明する。

【００３８】

本実施形態は、第１の実施形態と同じく、図1のブロック図で示される。すなわち、本実施形態は、テキスト文書を入力するテキスト入力部101と、入力されたテキスト文書を格納するテキスト記憶部102と、入力されたテキスト文書を生成したと推測される、テキスト文書の話題を隠れ変数に、テキスト文書を観測変数にそれぞれ対応付けた、単一もしくは複数のモデルを生成する仮モデル生成部103と、仮モデル生成部103が生成した各モデルを規定する各モデルパラメータの値を初期化するモデルパラメータ初期化部104と、モデルパラメータ初期化部104によって初期化されたモデルとテキスト記憶部102に格納されたテキスト文書を使ってモデルパラメータを推定するモデルパラメータ推定部105と、モデルパラメータ推定部105が行ったパラメータ推定の結果を格納する推定結果記憶部106と、推定結果記憶部106に複数のモデルのパラメータ推定結果が格納されている場合にその中から1つのモデルのパラメータ推定結果を選択するモデル選択部107と、モデル選択部107が選択したモデルのパラメータ推定結果から入力テキスト文書の分割を行って結果を出力するテキスト分割結果出力部108を備える。各々の部は、それぞれ計算機上に記憶されたプログラムによって、またはこのプログラムが記録された記録媒体を読み取ることによって動作させることにより実現可能である。

【００３９】

次に、本実施形態の動作について、順を追って説明する。

【００４０】
テキスト入力部101、テキスト記憶部102および仮モデル生成部103は、それぞれ先に述べた第１の実施形態におけるテキスト入力部101、テキスト記憶部102および仮モデル生成部103と同一の動作をする。テキスト記憶部102が入力テキスト文書を、単語の列、あるいは隣接する単語の2つ組、3つ組、もしくは一般のn個組の列として格納することや、入力テキスト文書に単語間スペースのない日本語の場合、公知の形態素解析法を適用することで、単語列として扱うことができることなども、第１の実施形態と同様である。

【００４１】
モデルパラメータ初期化部104は、仮モデル生成部103が生成したすべてのモデルについて、モデルを規定するパラメータの値を初期化する。モデルは、第１の実施形態と同様、left-to-right型離散HMMであるが、さらにタイドミクスチャ(tied-mixture)HMMであるとする。すなわち、状態iからの記号出力が、M個の記号出力確率$b_{1,j}$, $b_{2,j}$, …, $b_{M,j}$の線形結合$c_{i,1}b_{1,j}+c_{i,2}b_{2,j}+\cdots c_{i,M}b_{M,j}$であり、$b_{i,j}$の値は全状態にわたって共通とする。Mは一般には状態数Nよりも小さい、任意の自然数である。タイドミクスチャHMMについては、例えば文献「1995年11月、NTTアドバンステクノロジ株式会社、ラビナー他著、古井他訳、音声認識の基礎(下)、第280～281頁」（非特許文献７）に記載がある。タイドミクスチャ(tied-mixture)HMMのモデルパラメータは、状態遷移確率 a_i、全状態で共通の記号出力確率$b_{j,k}$、および記号出力確率に対する重み係数$c_{i,j}$である。ここで、i=1,2,…,Nで、Nは状態数である。j=1,2,…,Mで、Mは話題の種類数。また k=1, 2, …, L で、Lは入力テキスト文書に含まれる単語の種類数、すなわち語彙数である。状態遷移確率a_iは、第１の実施形態と同様、状態iから状態i+1に遷移する確率である。記号出力確率$b_{j,k}$は、話題jにおいて、インデクスkで指定される単語が出力される確率である。また重み係数$c_{i,j}$は、状態iにおいて話題jが発生する確率である。第１の実施形態と同様、記号出力確率の総和$b_{j,1}+b_{j,2}+\cdots+b_{j,L}$は1でなければならない。また、重み係数の総和$c_{i,1}+c_{i,2}+\cdots+c_{i,L}$も1でなければならない。

【００４２】
モデルパラメータ初期化部104は、状態数Nのモデルに対して、例えば上述の各パラメータの値を$a_i=N/T$、$b_{j,k}=1/L$、$c_{i,j}=1/M$のように設定する。この初期値の与え方に決まったやり方はなく、上述の確率の条件さえ満たしていれば、いろいろな方法があり得る。ここで述べた方法はほんの一例である。

【００４３】
モデルパラメータ推定部105は、モデルパラメータ初期化部104によって初期化された単一もしくは複数のモデルを順次受け取り、モデルが入力テキスト文書o_1, o_2, …, o_Tを生成する確率、すなわち尤度がなるべく高くなるように、モデルパラメータを推定する。これには、第１の実施形態と同様、期待値最大化法(EM法)を用いることができる。すなわち、その時点で得られているパラメータ値a_i、$b_{j,k}$、$c_{i,j}$を用いて、[数６]に示す漸化式によって前向き変数$\alpha_t(i)$および後向き変数$\beta_t(i)$をt=1, 2, …, T、i=1, 2, …, Nにわたって計算し、さらに[数７]に従ってパラメータ値を再計算する。再計算されたパラメータ値を用いて再度[数６]および[数７]を計算する。以下、収束するまで十分な回数これをくり返す。ただしここにδ_{ij}はクロネッカーのデルタ、すなわち、i=jなら1、そうでなければ0をとる。

【００４４】

【数６】

$$\alpha_1(i)=\sum_{j=1}^{M}c_{1,j}b_{j,o_1}\delta_{1,i},\ \alpha_t(i)=\sum_{j=1}^{M}\left\{a_{i-1}c_{i,j}b_{j,o_t}\alpha_{t-1}(i-1)+(1-a_i)c_{i,j}b_{j,o_t}\alpha_{t-1}(i)\right\},$$

$$\beta_T(i)=a_N\delta_{N,i},\ \beta_t(i)=\sum_{j=1}^{M}\left\{(1-a_i)c_{i,j}b_{j,o_{t+1}}\beta_{t+1}(i)+a_ic_{i+1,j}b_{j,o_{t+1}}\beta_{t+1}(i+1)\right\}.$$

【００４５】

【数７】

$$a_i\leftarrow\frac{\sum_{t=1}^{T-1}\sum_{j=1}^{M}\alpha_t(i)a_ic_{i+1,j}b_{j,o_t}\beta_{t+1}(i+1)}{\sum_{t=1}^{T-1}\sum_{j=1}^{M}\left\{\alpha_t(i)(1-a_i)c_{i,j}b_{j,o_t}\beta_{t+1}(i)+\alpha_t(i)a_ic_{i+1,j}b_{j,o_t}\beta_{t+1}(i+1)\right\}},$$

$$b_{ij}\leftarrow\frac{\sum_{t=1}^{T}\sum_{i=1}^{N}\left\{\alpha_t(i)(1-a_i)c_{i,j}b_{j,o_t}\beta_{t+1}(i)+\alpha_t(i)a_ic_{i+1,j}b_{j,o_t}\beta_{t+1}(i+1)\right\}}{\sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{k=1}^{L}\left\{\alpha_t(i')(1-a_{i'})c_{i',j}b_{j,k}\beta_{t+1}(i')+\alpha_t(i')a_{i'}c_{i'+1,j}b_{j,k}\beta_{t+1}(i'+1)\right\}},$$

$$c_{ij}\leftarrow\frac{\sum_{t=1}^{T}\left\{\alpha_t(i)(1-a_i)c_{i,j}b_{j,o_t}\beta_{t+1}(i)+\alpha_t(i)a_ic_{i+1,j}b_{j,o_t}\beta_{t+1}(i+1)\right\}}{\sum_{j'=1}^{M}\sum_{t=1}^{T}\left\{\alpha_t(i)(1-a_i)c_{i,j'}b_{j',o_t}\beta_{t+1}(i)+\alpha_t(i)a_ic_{i+1,j'}b_{j',o_t}\beta_{t+1}(i+1)\right\}}.$$

【００４６】

モデルパラメータ推定部105におけるパラメータ推定の反復計算の収束判定を行うには、尤度の上昇量をみればよい。すなわち、上述の反復計算によって尤度の上昇がみられなくなれば、その時点で反復計算を終了すればよい。ここに、尤度は$\alpha_1(1)\beta_1(1)$として得られる。モデルパラメータ推定部105は、反復計算を終了した時点で、モデルパラメータa_i、$b_{j,k}$、$c_{i,j}$と、前向きおよび後向き変数$\alpha_t(i)$、$\beta_t(i)$を、モデル(HMM)の状態数と対にして、推定結果記憶部106に格納する。

【００４７】

モデル選択部107は、第１の実施形態と同様、モデルパラメータ推定部105で状態数ごとに得られたパラメータ推定結果を推定結果記憶部106から受け取り、各モデルの確からしさを計算し、もっとも確からしいモデルを1つ選択する。モデルの確からしさは、公知の赤池情報量基準(AIC)や最小記述長基準(MDL基準)などに基づいて計算することができる。また、第１の実施形態と同様、AICでもMDLでも、モデルパラメータ数NLに関わる項に、経験的に決まる定数係数をかけて、選択されるモデルを意図的に調整する操作も行って差し支えない。

【００４８】

テキスト分割結果出力部108は、第１の実施形態におけるテキスト分割結果出力部108と同様、モデル選択部107によって選択された状態数すなわち話題数Nのモデルに対応するモデルパラメータ推定結果を推定結果記憶部106から受け取り、この推定結果における入力テキスト文書に対する話題ごとの分割結果を算出する。最終的な分割結果は、[数４]に従

って、$P(z_t=i \mid o_1, o_2, \cdots, o_T)$ が最大となるiをt=1，2，…，T にわたって求めることで得られる。

【００４９】

なお、モデルパラメータ推定部105は、第１の実施形態と同様、最尤推定法の代わりに最大事後確率推定(MAP推定)法によってモデルパラメータを推定してもよい。

【００５０】

（第３の実施形態）

次に、本発明の第３の実施形態について、図面を参照して説明する。

【００５１】

本実施形態は、第１および第２の実施形態の例と同じく、図1のブロック図で示される。すなわち、本実施形態は、テキスト文書を入力するテキスト入力部101と、入力されたテキスト文書を格納するテキスト記憶部102と、入力されたテキスト文書を生成したと推測される、テキスト文書の話題を隠れ変数に、テキスト文書を観測変数にそれぞれ対応付けた、単一もしくは複数のモデルを生成する仮モデル生成部103と、仮モデル生成部103が生成した各モデルを規定する各モデルパラメータの値を初期化するモデルパラメータ初期化部104と、モデルパラメータ初期化部104によって初期化されたモデルとテキスト記憶部102に格納されたテキスト文書を使ってモデルパラメータを推定するモデルパラメータ推定部105と、モデルパラメータ推定部105が行ったパラメータ推定の結果を格納する推定結果記憶部106と、推定結果記憶部106に複数のモデルのパラメータ推定結果が格納されている場合にその中から1つのモデルのパラメータ推定結果を選択するモデル選択部107と、モデル選択部107が選択したモデルのパラメータ推定結果から入力テキスト文書の分割を行って結果を出力するテキスト分割結果出力部108を備える。各々の部は、それぞれ計算機上に記憶されたプログラムによって、またはこのプログラムが記録された記録媒体を読み取ることによって動作させることにより実現可能である。

【００５２】

次に、本実施形態の動作について、順を追って説明する。

【００５３】

テキスト入力部101、テキスト記憶部102および仮モデル生成部103は、それぞれ先に述べた第１および第２の実施形態におけるテキスト入力部101、テキスト記憶部102および仮モデル生成部103と同一の動作をする。テキスト記憶部102が入力テキスト文書を、単語の列、あるいは隣接する単語の2つ組、3つ組、もしくは一般のn個組の列として格納することや、入力テキスト文書に単語間スペースのない日本語の場合、公知の形態素解析法を適用することで、単語列として扱うことができることなども、本発明の第１および第２の実施形態と同様である。

【００５４】

モデルパラメータ初期化部104は、仮モデル生成部103が生成した単一または複数のモデル各々について、モデルパラメータ、すなわち状態遷移確率a_iおよび記号出力確率b_{ij}を確率変数として、ある種の分布を仮定し、それらの分布を規定するパラメータの値を初期化する。以下では、モデルパラメータの分布を規定するパラメータを、元のパラメータに対してメタパラメータと呼ぶことにする。つまり、モデルパラメータ初期化部104はメタパラメータの初期化を行う。本実施形態では、状態遷移確率a_iおよび記号出力確率b_{ij}の分布として、それぞれベータ分布 $\log p(a_i \mid \kappa_{0,i}, \kappa_{1,i}) = (\kappa_{0,i}-1)\times\log(1-a_i) + (\kappa_{1,i}-1)\times\log(a_i) + \text{const}$、ディレクレ分布 $\log p(b_{i,1}, b_{i,2}, \cdots, b_{i,L} \mid \lambda_{i,1}, \lambda_{i,2}, \cdots, \lambda_{i,L}) = (\lambda_{i,1}-1)\times\log(b_{i,1}) + (\lambda_{i,2}-1)\times\log(b_{i,2}) + \cdots + (\lambda_{i,L}-1)\times\log(b_{i,L}) + \text{const}$を使用する。メタパラメータは$\kappa_{0,i}$，$\kappa_{1,i}$，$\lambda_{i,j}$である。ここで、i=1,2,…,N、j=1,2,…,Lである。モデルパラメータ初期化部104は、例えば$\kappa_{0,i}=\kappa_0$，$\kappa_{1,i}=\kappa_1$，$\lambda_{ij}=\lambda_0$，ただし$\kappa_0=\varepsilon(1-N/T)+1$，$\kappa_1=\varepsilon N/T+1$，$\lambda_0=\varepsilon/L+1$、というようにメタパラメータを初期化する。$\varepsilon$としては、0.01などのように適当な正数を当てる。なお、初期値の与え方に決まったやり方はなく、いろいろな方法があり得る。この初期化方法はほんの一例である。

【００５５】

モデルパラメータ推定部105は、モデルパラメータ初期化部104によって初期化された単一もしくは複数のモデルを順次受け取り、モデルが入力テキスト文書o_1，o_2，…，o_Tを生成する確率、すなわち尤度がなるべく高くなるように、メタパラメータを推定する。これにはベイズ推定法から導出される公知の変分ベイズ法を用いることができる。すなわち、例えば文献「2002年７月、電子情報通信学会誌、第85巻、第7号、第504～509頁、上田、ベイズ学習〔III〕 —変分ベイズ学習の基礎—」（非特許文献８）に記載があるように、その時点で得られているメタパラメータ値$\kappa_{0,i}$，$\kappa_{1,i}$，$\lambda_{i,j}$を用いて、[数８]に示す漸化式によって前向き変数$\alpha_t(i)$および後向き変数$\beta_t(i)$をt=1，2，…，T、i=1，2，…，Nにわたって計算し、さらに[数９]に従ってメタパラメータ値を再計算する。再計算されたパラメータ値を用いて、再度[数８]および[数９]を計算する。以下、収束するまで十分な回数これをくり返す。ただしここに、δ_{ij}はクロネッカーのデルタ、すなわち、i=jなら1、そうでなければ0をとる。また、$\Psi(x)=d(\log\Gamma(x))/dx$で、$\Gamma(x)$はガンマ関数である。

【００５６】

【数８】

$$\alpha_1(i)=\exp(B_{i,o_1})\delta_{1,i},$$
$$\alpha_t(i)=\alpha_{t-1}(i-1)\exp(A_{1,i-1}+B_{i,o_t})+\alpha_{t-1}(i)\exp(A_{0,i}+B_{i,o_t}),$$
$$\beta_T(i)=\exp(A_{1,N})\delta_{N,i},$$
$$\beta_t(i)=\beta_{t+1}(i)\exp(A_{0,i}+B_{i,o_{t+1}})+\beta_{t+1}(i+1)\exp(A_{1,i}+B_{i+1,o_{t+1}}).$$

ただし

$$A_{0,i}=\Psi(\kappa_{0,i})-\Psi(\kappa_{0,i}+\kappa_{1,i}),$$
$$A_{1,i}=\Psi(\kappa_{1,i})-\Psi(\kappa_{0,i}+\kappa_{1,i}),$$
$$B_{ik}=\Psi(\lambda_{ik})-\Psi\left(\sum_{j=1}^{L}\lambda_{ij}\right).$$

【００５７】

【数９】

$$\kappa_{0,i} \leftarrow \kappa_0 + \sum_{t=1}^{T-1} \overline{z_{t,i} z_{t+1,i}},\quad \kappa_{1,i} \leftarrow \kappa_1 + \sum_{t=1}^{T-1} \overline{z_{t,i} z_{t+1,i+1}} + \delta_{N,i},\quad \lambda_{ik} \leftarrow \lambda_0 + \sum_{t=1}^{T-1} \overline{z_{t,i}} \delta_{k,o_t}.$$

ただし

$$\overline{z_{t,i}} = \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{N} \alpha_t(j)\beta_t(j)},$$

$$\overline{z_{t,i} z_{t+1,i}} = \frac{\alpha_t(i)\exp\left(A_{0,i} + B_{i,o_{t+1}}\right)\beta_{t+1}(i)}{\sum_{j=1}^{N}\sum_{s=\{0,1\}} \alpha_t(j)\exp\left(A_{s,j} + B_{j+s,o_{t+1}}\right)\beta_{t+1}(j+s)},$$

$$\overline{z_{t,i} z_{t+1,i+1}} = \frac{\alpha_t(i)\exp\left(A_{1,i} + B_{i+1,o_{t+1}}\right)\beta_{t+1}(i+1)}{\sum_{j=1}^{N}\sum_{s=\{0,1\}} \alpha_t(j)\exp\left(A_{s,j} + B_{j+s,o_{t+1}}\right)\beta_{t+1}(j+s)}.$$

【００５８】

モデルパラメータ推定部105におけるパラメータ推定の反復計算の収束判定は、近似的尤度の上昇量をみればよい。すなわち、上述の反復計算によって近似的尤度の上昇がみられなくなれば、その時点で反復計算を終了すればよい。ここで、近似的尤度とは、前向き変数と後向き変数の積 $\alpha_1(1)\beta_1(1)$ として得られる。モデルパラメータ推定部105は、反復計算を終了した時点で、メタパラメータ $\kappa_{0,i}$、$\kappa_{1,i}$、$\lambda_{i,j}$ と、前向きおよび後向き変数 $\alpha_t(i)$、$\beta_t(i)$ を、モデル(HMM)の状態数Nと対にして、推定結果記憶部106に格納する。

【００５９】

なお、モデルパラメータ推定部105におけるメタパラメータのベイズ推定法としては、上述の変分ベイズ法以外にも、公知のマルコフ連鎖モンテカルロ法やラプラス近似法など、任意の方法を使うことができる。本実施形態は、変分ベイズ法に限定されるものではない。

【００６０】

モデル選択部107は、モデルパラメータ推定部105で状態数ごとに得られたパラメータ推定結果を推定結果記憶部106から受け取り、各モデルの確からしさを計算し、もっとも確からしいモデルを1つ選択する。モデルの確からしさは、例えば上述した変分ベイズ法の枠組みでは、公知のベイズ的基準を使用することができる。ベイズ的基準は[数１０]で計算可能である。[数１０]において P(N) は状態数すなわち話題数Nの事前確率で、あらかじめ何らかの方法で定めておく。取り立てて理由がなければ、P(N) は一定値でよい。逆に、特定の状態数が起こりやすい、あるいは起こりにくいということが事前にわかっている場合は、特定の状態数に対応するP(N)を大きく、あるいは小さく設定する。また、[数１０]に現れるメタパラメータ $\kappa_{0,i}$、$\kappa_{1,i}$、$\lambda_{i,j}$ と、前向きおよび後向き変数 $\alpha_t(i)$、$\beta_t(i)$ としては、状態数Nに対応するものを推定結果記憶部106から取得して用いる。

【００６１】

【数１０】

$$P(N)\alpha_1(1)\beta_1(1)$$
$$\times\exp\left\{\sum_{i=1}^{N}(\kappa_{0,i}-\kappa_0)(\Psi(\kappa_{0,i}+\kappa_{1,i})-\Psi(\kappa_{0,i}))+\sum_{i=1}^{N}(\kappa_{1,i}-\kappa_1)(\Psi(\kappa_{0,i}+\kappa_{1,i})-\Psi(\kappa_{1,i}))\right\}$$
$$\times\exp\left\{\sum_{i=1}^{N}\sum_{k=1}^{L}(\lambda_{ij}-\lambda_0)\left(\Psi\left(\sum_{j=1}^{L}\lambda_{ij}\right)-\Psi(\lambda_{ik})\right)\right\}$$
$$\times\prod_{i=1}^{N}\left\{\frac{\Gamma(\kappa_0+\kappa_1)}{\Gamma(\kappa_{0,i}+\kappa_{1,i})}\frac{\Gamma(\kappa_{0,i})\Gamma(\kappa_{1,i})}{\Gamma(\kappa_0)\Gamma(\kappa_1)}\frac{\Gamma\left(\sum_{j=1}^{L}\lambda_0\right)}{\Gamma\left(\sum_{j=1}^{L}\lambda_{ij}\right)}\prod_{j=1}^{L}\frac{\Gamma(\lambda_{ij})}{\Gamma(\lambda_0)}\right\}$$

【００６２】

テキスト分割結果出力部108は、上述の第１および第２の実施形態におけるテキスト分割結果出力部108と同様、モデル選択部107によって選択された状態数すなわち話題数Nのモデルに対応するモデルパラメータ推定結果を推定結果記憶部106から受け取り、この推定結果における入力テキスト文書に対する話題ごとの分割結果を算出する。最終的な分割結果は、[数４]に従って、$P(z_t=i \mid o_1, o_2, \cdots, o_T)$ が最大となるiをt=1, 2, …, Tにわたって求めることで得られる。

【００６３】

なお、本実施形態でも、上述した第２の実施形態と同様、通常のleft-to-right型HMMの代わりに、タイドミクスチャ(tied-mixture)型のleft-to-right型HMMを生成、初期化、パラメータ推定するように、仮モデル生成部103、モデルパラメータ初期化部104、モデルパラメータ推定部105をそれぞれ構成することが可能である。

【００６４】

（第４の実施形態）

次に、本発明の第４の実施形態について、図面を参照して詳細に説明する。

【００６５】

図4を参照すると、本発明の第４の実施形態は、テキスト処理プログラムを記録した記録媒体601を備える。この記録媒体601はCD-ROM、磁気ディスク、半導体メモリその他の記録媒体であってよく、ネットワークを介して流通する場合も含む。テキスト処理プログラムは記録媒体601からデータ処理装置602に読み込まれ、データ処理装置602の動作を制御する。

【００６６】

本実施形態としては、データ処理装置602はテキスト処理プログラムの制御により、第１、第２、もしくは第３の実施形態におけるテキスト入力部101、仮モデル生成部103、モデルパラメータ初期化部104、モデルパラメータ推定部105、モデル選択部107、テキスト分割結果出力部108による処理と同一の処理を実行して、第１、第２、もしくは第３の実施形態におけるテキスト記憶部102、推定結果記憶部106とそれぞれ同等の情報を有するテキスト記録媒体603、モデルパラメータ推定結果記録媒体604を参照することによって、入力されたテキスト文書に対する話題ごとの分割結果を出力する。

【図面の簡単な説明】

【００６７】

【図１】本発明のテキスト処理装置の構成を示したブロック図である。

【図２】本発明のテキスト処理装置の動作を説明するためのフローチャートである。

【図３】隠れマルコフモデルを説明するための概念図である。

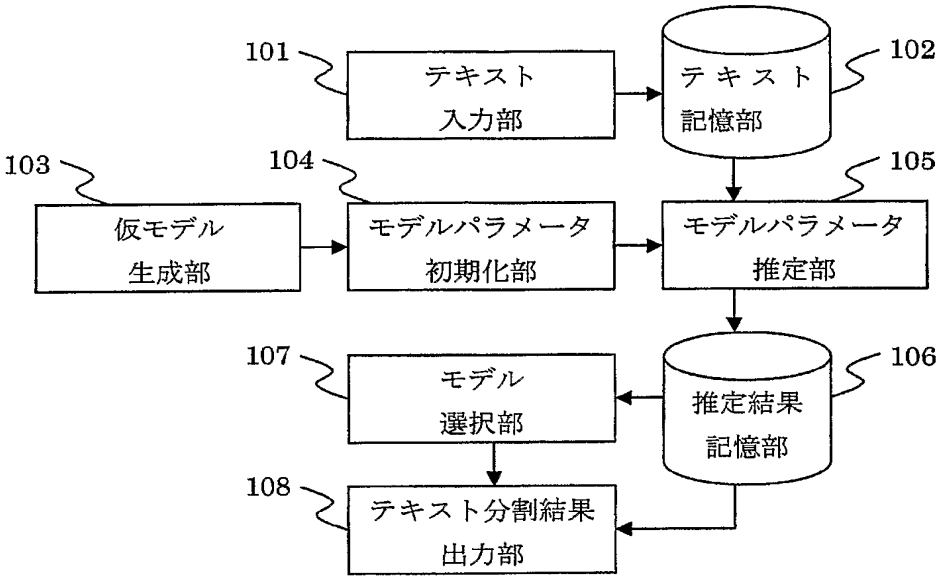
【図４】本発明のテキスト処理装置の構成の他の態様を示したブロック図である。
【図５】第１の従来技術を説明するための概念図である。
【図６】第２の従来技術を説明するための概念図である。

【符号の説明】

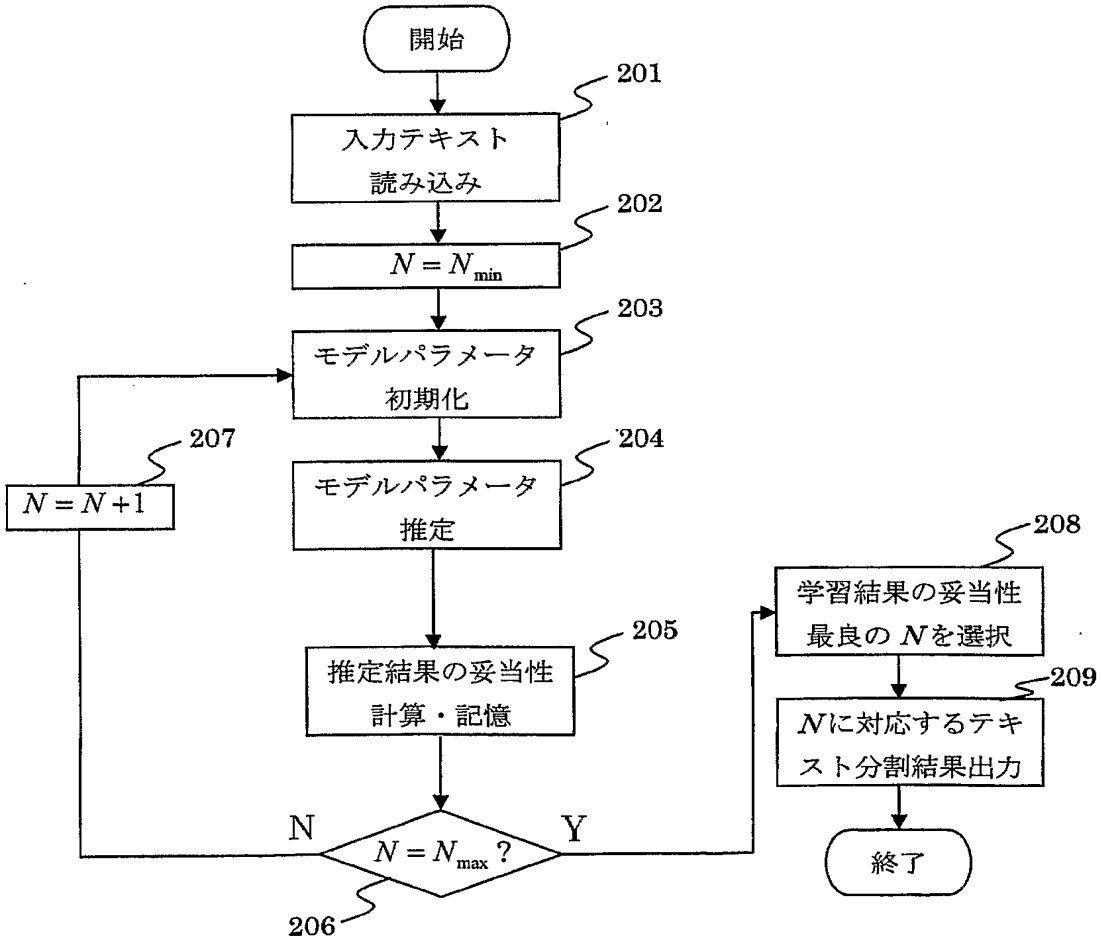
【００６８】

101　テキスト入力部
102　テキスト記憶部
103　仮モデル生成部
104　モデルパラメータ初期化部
105　モデルパラメータ推定部
106　推定結果記憶部
107　モデル選択部
108　テキスト分割結果出力部
201　入力テキスト読み込み
202　状態数初期化 $N=N_{min}$
203　モデルパラメータ初期化
204　モデルパラメータ推定
205　推定結果の妥当性計算・記憶
206　条件分岐 $N=N_{max}$?
207　状態数変更 N=N+1
208　学習結果の妥当性最良のNを選択
209　Nに対応するテキスト分割結果出力
601　記録媒体
602　データ処理装置
603　テキスト記録媒体
604　モデルパラメータ推定結果記録媒体

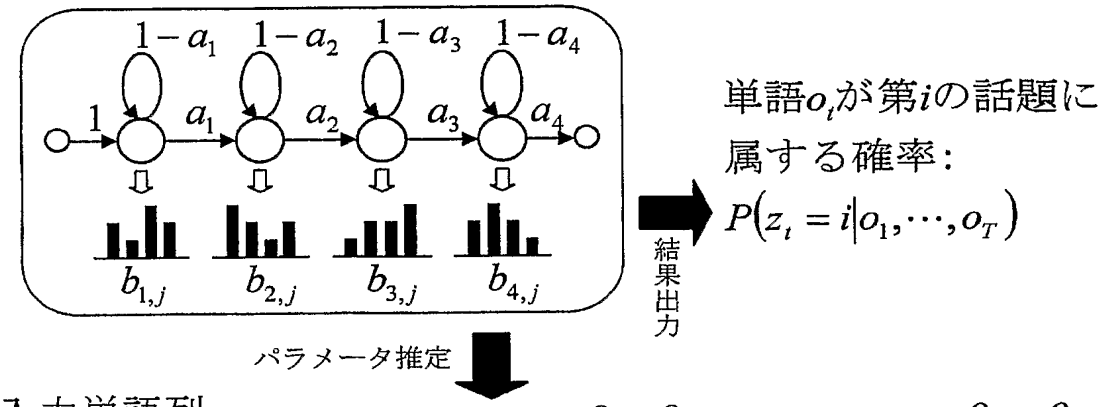
【書類名】図面
【図１】


101
テキスト
入力部
102
テキスト
記憶部
103
仮モデル
生成部
104
モデルパラメータ
初期化部
105
モデルパラメータ
推定部
107
モデル
選択部
106
推定結果
記憶部
108
テキスト分割結果
出力部

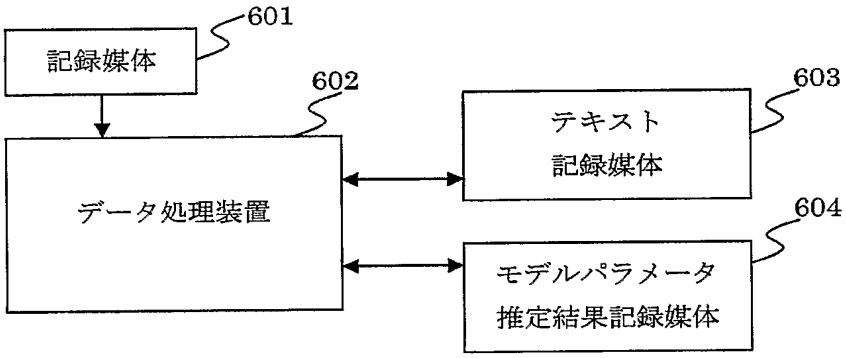
【図２】


開始
201
入力テキスト
読み込み
202
$N = N_{min}$
203
モデルパラメータ
初期化
204
モデルパラメータ
推定
205
推定結果の妥当性
計算・記憶
206
$N = N_{max}$?
N
Y
207
$N = N + 1$
208
学習結果の妥当性
最良の N を選択
209
N に対応するテキ
スト分割結果出力
終了

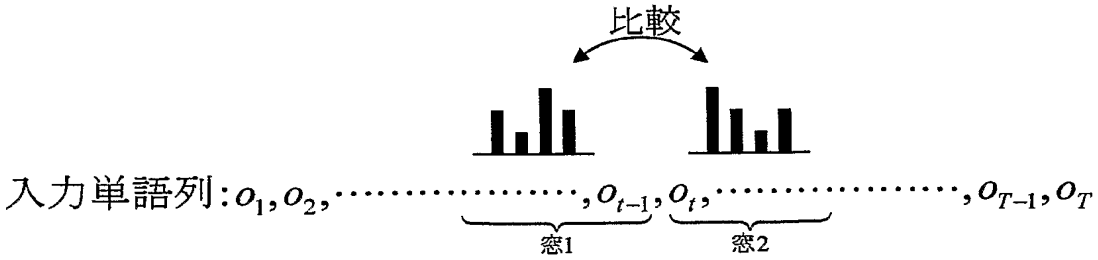
【図３】


1−a1
1−a2
1−a3
1−a4
1
a1
a2
a3
a4
b1,j
b2,j
b3,j
b4,j
結果出力
単語ot が第iの話題に
属する確率：
P(zt = i|o1,…,oT)
パラメータ推定
入力単語列：o1,o2,……,ot−1,ot,……,oT−1,oT

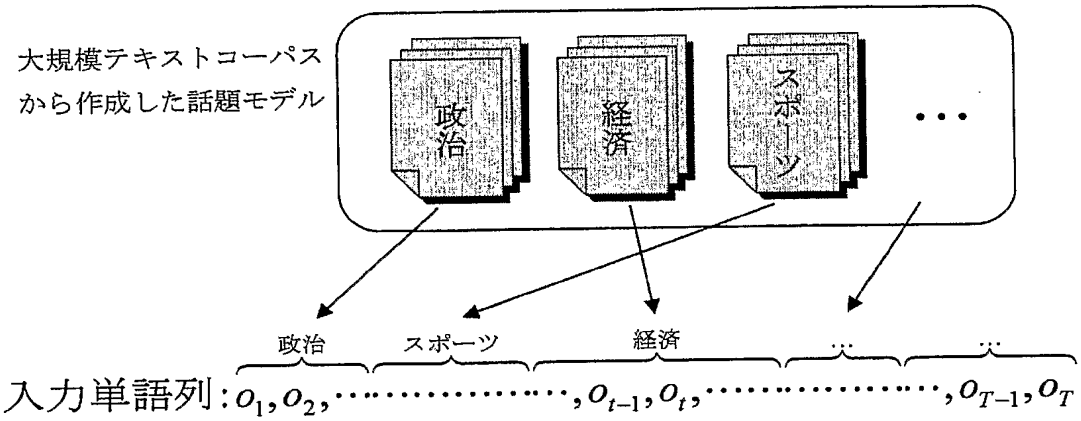
【図４】


601
記録媒体
602
データ処理装置
603
テキスト
記録媒体
604
モデルパラメータ
推定結果記録媒体

【図５】


比較
入力単語列：o1,o2,……,ot−1,ot,……,oT−1,oT
窓1
窓2

【図６】


大規模テキストコーパス
から作成した話題モデル
政治
経済
スポーツ
・・・
政治
スポーツ
経済
入力単語列：$o_1, o_2, \cdots, o_{t-1}, o_t, \cdots, o_{T-1}, o_T$

【書類名】要約書
【課題】　入力テキスト文書の性質によってパラメータを調整する手間が不要で、事前に時間と費用をかけて大規模なテキストコーパスを準備する必要もなく、なおかつ入力テキストのドメインに依存せずにテキストを話題ごとに分割できるテキスト処理装置を提供する。
【解決手段】　入力されたテキスト文書を生成したと推測されるテキスト文書の話題を隠れ変数に、テキスト文書を観測変数にそれぞれ対応付けた、１つまたは複数の確率モデルを生成する仮モデル生成部103と、入力されたテキスト文書を用いて、確率モデルを規定するパラメータを推定するモデルパラメータ推定部105を有する。パラメータが推定された後に、モデル選択部107が最良の確率モデルを選択し、テキスト分割結果出力部108が、その最良の確率モデルに対応する推定結果を用いて、入力テキスト文書中の話題の推移を確率的に算出する。
【選択図】　図１

特願２００４－００９１４４

出　願　人　履　歴　情　報

識別番号　［０００００４２３７］

１．変更年月日　１９９０年　８月２９日
［変更理由］　新規登録
住　所　東京都港区芝五丁目７番１号
氏　名　日本電気株式会社